A New Storm Topology for Synopsis Management in the Processing Architecture

Abstract— The Processing Architecture based on Measurement Metadata (PAbMM) is a data stream management system specialized in measurement and evaluation (M&E) projects, which incorporates predictive and detective behavior on data streams. It uses a case based organizational memory for recommending courses of action in each detected online situation and previously modeled by the project definition. In this work the storm topology associated with the online processing in PAbMM is described. Additionally, a new synopses strategy for monitoring entities under analysis is presented and a new schema for training the online classifiers is introduced. This new schema allows indicating to the classifiers the problem characterization, the proposal solution and the associated indicator value (target class). A practical case associated with the weather radar of the Experimental Agricultural Station (EAS) INTA Anguil (Province of La Pampa, Argentina) is shown, indicating the advantages of this storm topology and the new schema oriented to training data set.

Index Terms—Data Stream Mining, Data Streams, Measurement, Evaluation, Stream Management

1 Introduction

Nowadays, the real time monitoring has gained a particular importance in different kind of software applications and devices[1]. This current context is characterized between others by the Wireless Sensor Networks (WSN) which can be defined as a network of tiny devices, spatially distributed and work cooperatively to comunicate data gathered from the monitored field through wireless links. The sensors or tiny devices are likely heterogeneous and their processing capacities are limited to the communication and the sending of raw data associated with the measurement[2].

The Processing Architecture based on Measurement Metadata (PAbMM) is a data stream management system specialized in measurement and evaluation projects. It is supported by an M&E framework called C-INCAMI (*Context-Information Need, Concept Model, Attribute, Metric and Indicator*)[3],[4] which fosters the repeatability, comparability, extensibility and consistence associated with the measurement process. Each M&E project is defined through of GOCAME (*Goal Oriented Context-Aware Measurement and Evaluation*) strategy[5], using the concepts and relationships defined in C-INCAMI. From each M&E project definition (PD), PAbMM can incorporate heterogeneous data sources for implemmenting each metric.

The data streams coming from the heterogeneous data sources (i.e. sensors) are organized by C-INCAMI/MIS 2 (*Measurement Interchange Schema, version 2*)[6], which allows using data and metadata jointly for guiding the online processing strategy in PAbMM. The online processing strategy was described in [7] using SPEM (*Software and Systems Process Engineering Metamodel*)[8] and was updated in [9] for supporting a case based organizational memory.

The online processing strategy associated with PAbMM uses Apache Storm[10], while the big data repository related to the the initial training dataset for the online classifiers and the M&E project definition utilizes Apache HBase[11].

The main contributions of this work are as follows: a) a new schema called CINCAMI/TS (*Training Set*) is introduced with the aim to guides the online classifiers training process. It is important because now it is possible not just the using the training data set, but also the using of the learned experience from the organizational memory associating problem with estimated solution, b) the Apache Storm topology associated with the online processing in PAbMM is refined and detailed in relation with its formal definition in [7],[9] and c) A new synopses strategy is incorporated for improving the last known state of each entity under analysis (i.e. online monitoring).

This article is organized in seven sections. Section 2 outlines the own updating of C-INCAMI and the ontology supporting the Organizational Memory. Section 3 reviews the C-INCAMI/MIS version 2 and introduces the new schema called C-INCAMI/TS. Section 4 synthesizes the PAbMM conceptual architecture, introduces its associated storm topology and the synopses management. Section 5 summarizes the application associated with the Weather Radar of the EAS INTA Anguil (La Pampa, Argentina). Section 6 discusses related works and finally the conclusions are presented.

2 C-INCAMI and The Organizational Memory

Originally, the C-INCAMI framework was oriented for definiting and implementing the evaluation and measurement processes inside software organizations. It defined the neccesary concepts and relationshiphs guided

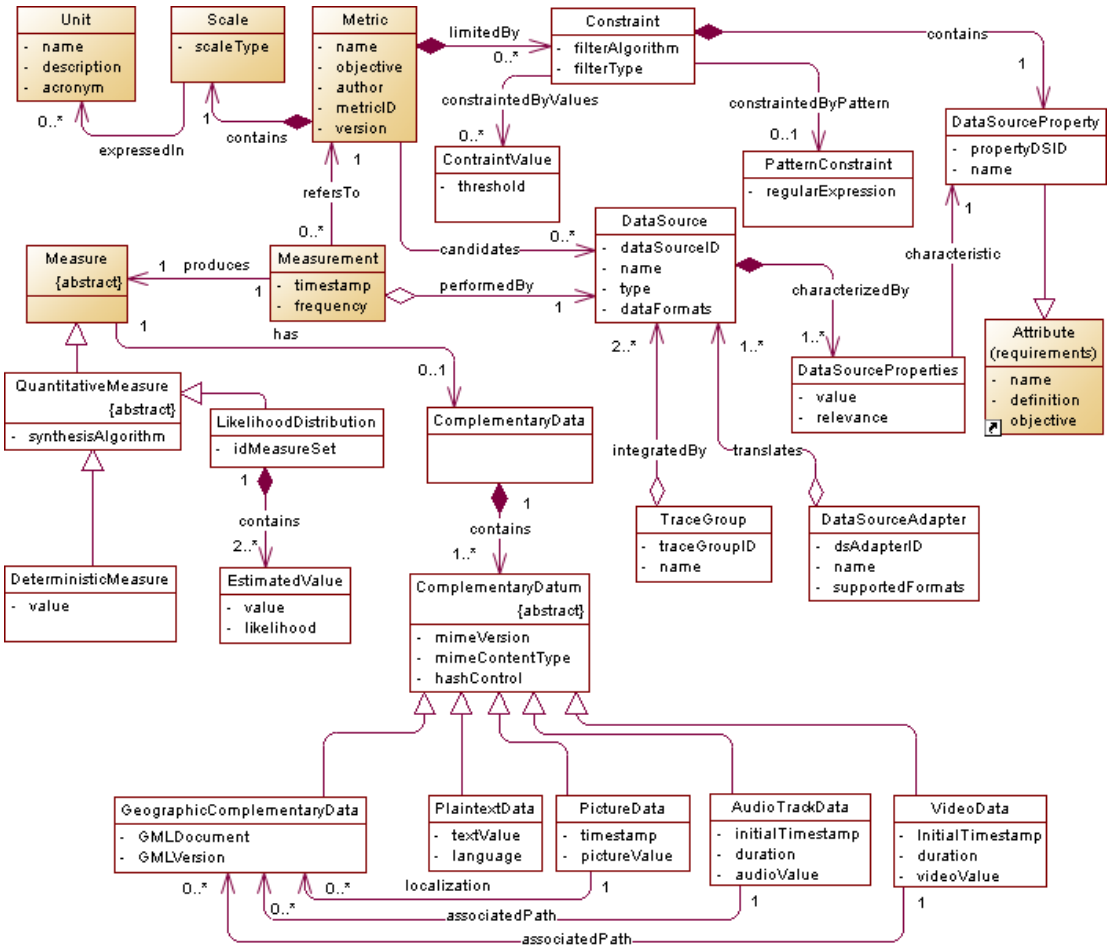

Fig. 1. Extensions for the measurement component of C-INCAMI.

by an information need associated with an entity under and evaluation is applicable to many areas, but some of them need to extend or complement the associated concepts and their relationships for better describing of the entity under analysis[5],[12].

In relation with PAbMM, the Organizational Memory (OM) allows capitalize the learned experiences and using them for recommending courses of action in the decision-making process. In this sense, it is convenient the case-based organization because the solution and problem may be characterized by means of attributes. Because each metric is associated with an entity's attribute in C-INCAMI, its associated measurement could be considered as the quantification of the problem's characteristic. Thus, it could be possible to establish a relation between the entity's attribute and problem's characteristic with the aim of improving the recommendation strategy.

The section 2.1 synthesizes the extentions on the C-INCAMI framework for better supporting the data generated from the Weather Radar of INTA Anguil. The section 2.2 outlines the case based organizational memory.

2.1 The C-INCAMI Extensions

The C-INCAMI is conceptual framework [3],[4], which defines the concepts and their related components for Measurement and Evaluation (M&E) area in software organizations. It provides a domain (ontological) model defining all the terms, properties and relationships needed for designing and implementing the M&E processes. It is an approach in which the requirements specification, M&E, and analysis of results are performed for satisfying a specific information need, in a given context. In C-INCAMI, concepts and relationships are intended to be used along all the M&E activities. This way, a common understanding of data and metadata is shared among projects fostering a more consistent analysis.

Figure 1 shows the extension for the measurement package of C-INCAMI, however the conceptual framework is structured in six components, namely: i) *M&E project*, ii) *Non-functional Requirements*, iii) *Context*, iv) *Measurement*, v) *Evaluation*, and vi) *Analysis and Recommendation*. The *M&E Project Definition* component defines and relates a set of project needed for dealing with M&E activities, methods, roles and artifacts. The *non-functional requirements component* allows specifying the information need of any M&E project. For the *context* package, one concept is *Context*, which represents the relevant state of the situation of the entity to be assessed with regard to the information need. C-INCAMI considers the Context as a special kind of *Entity* in which related relevant entities are involved. For context description, attributes of the relevant entities are used – which are also Attributes called *Context Properties* (See [4] for details). The *Measurement* component includes the concepts and relationships intended for specifying the meas-

urement design and implementation. The *Evaluation* component includes the concepts and relationships intended to specify the evaluation design and implementation. It is worthy to mention that the selected metrics are useful for a measurement tasks as long as the selected indicators are useful for the evaluation tasks in order to interpret the stated information need. Finally, the *Analysis and Recommendation* Component establishes the feedback mechanism for supporting the M&E project.

For the Weather Radar (WR) of INTA [12], it was seen that the WR could provide measures, a picture or even picture sequences. In this context, the state of C-INCAMI was not expressive enough for describing the situation and for this reason the extensions shown in figure 1 were necessary[6]. For example, C-INCAMI considers the value associated with the *measure* just like a deterministic number excluding the possibility for modelling likelihood distributions. In figure 1, the extension of the concept called *measure* and the use of complementary data as a new point of view over each measure was proposed. For better differentiation, the original concepts from C-INCAMI were maintained with painted background in opossition with blank backgrounds for the new incorporated concepts or extensions.

As shown in figure 1, the *QuantitativeMeasure* class inherits from the *Measure* class and it represents the situation in which the measure is exclusively numeric, but not necessarily deterministic. In this sense, the *QuantitativeMeasure* class could be associated with a deterministic measurement process in which we get a clearly defined value (*DeterministicMeasure* class in figure 1), or it could be associated with an estimated process in which we obtain a set of <value, likelihood> pairs (See the classes called *LikelihoodDistribution* and *EstimatedValue* in figure 1). For this reason, the *QuantitativeMeasure* class incorporates the idea of *synthesisAlgorithm* as attribute, because even when is trivial in a deterministic value (the value is the same); it does not trivial in likelihood distribution in which we could use the mathematical expectation for synthesizing the distribution in one representative value.

Each measure could have complementary data (See the classes *ComplementaryData*) which allow complementing the measure itself, for example, one picture could describe the habitat under analysis when we want to measure the wind velocity of the place. So, the complementary data are a composition of at least one complementary datum, which is represented as an abstract class called *ComplementaryDatum* in the figure 1 and it incorporates three attributes: 1) *mimeVersion*: the mime version associated with the complementary datum, 2) *mimeContentType*: it identifies the type of content (e.g. image/jpeg, audio/mpeg, video/3gpp, etc.), and 3) *hashControl*: it is a footprint for verifying the integrity of the content (e.g. MD5 footprint). As you can see in figure 1, from the *ComplementaryDatum* class inherit five classes:

1. *GeographicComplementaryData*: it allows sending a document under the Geography Markup Language (GML)[13],[14] as a complement of the measure. This class has two attributes, *GMLDocument* and *GMLVersion*. The *GMLDocument* attribute contains the document organized in terms of GML and the *GMLVersion* attribute refers to the GML's version. It is worthy to mention because it allows incorporating the possibility for establishing a relationship between the positioning and the image, the audio track or the video;
2. *PlainTextData*: It allows incorporating textual information associated with the measure or the measurement device, for example, the device's log at the moment of the measure. This class incorporates the attributes *textValue* and *language*. The *textValue* attribute contains the data in text plain and the *language* attribute indicates the idiom in which the text is written by the use of ISO 639[15];
3. *PictureData*: This class incorporates the possibility of using a photo as complement of a measure. The *timestamp* attribute indicates the moment in which the picture was taken, and the *pictureValue* attribute store the data associated with the image itself. As you can see in figure 1, one *PictureData*'s object could be associated with positional data through the association with the class *GeographicComplementary*;
4. *AudioTrackData*: the class allows using an audio track as a complement of the measure. The *timestamp* attribute indicates the moment in which the audio track started the recording. The *duration* attribute is associated with the track's longitude and the *audioValue* attribute represents the audio itself. In this case, it is possible establishing a relationship between an *AudioTrackData*'s object and a *GeographicComplementaryData*'s object considering the timestamp. That is to say, if it can be synchronized the timestamp associated with the starting of the audio track with a described instant in the GML data, then it's highly possible analyze the audio and geographic information jointly.
5. *VideoData*: it represents the possibility of using a video as a complement of the measure. The *timestamp* attribute indicates the moment in which the video started the recording. The *duration* attribute is associated with the video's longitude and the *videoValue* attribute represents the video itself. Like the audio track, a relationship can be established between a *VideoData*'s object and a *GeographicComplementaryData*'s object considering the timestamp.

In this sense and trough the incorporation of the classes associated with the complementary data, it was possible to manage a new perspective of data for complementing each measure and its processing[6]. The C-INCAMI framework was extended for supporting complementary data in metrics associated with entity attributes and/or context properties. For taking advantage of the complementary data in the measurement processes, it was neccesary incorporates in C-INCAMI the specific figure of the data source because each measurement project has different requirements and the data collector is essential in terms of reliability, precision, etc.

With the aim of focusing the concept associated with the data collector, the *dataCollectorName* attribute from the *Measurement* class was replaced for a new *DataSource* class (See in figure 1). The *DataSource* class represents the concept associated with the measurement device, which allows getting the measures. In this sense, the *DataSource* class incorporates four representative attributes: *dataSourceID*, *name*, *type* and *dataFormats*. The *dataSourceID* attribute identifies the data source along the M&E projects for fostering the traceability. The *name* attribute is a more friendly way for referencing the device (e.g. an alias). The *type* attribute allows knowing if the device sends the measures in predictable or unpredictable way (e.g. the weather radar regularly sends data for processing. So, it is predictable). The *dataFormats* attribute allows knowing about the different ways that the data source could organize the content. Each data source sends the measures always through a measurement adapter (See *DataSourceAdapter* in figure 1). The measurement adapter translates from the original data format associated with the data source to the C-INCAMI/MIS stream (See Section 3). In this way, each data source adapter incorporates three attributes: *dsAdapterID*, *name* and *supportedFormats*. The *dsAdaptedID* attribute identifies the measurement adapter along the M&E projects. The *name* attribute is a friendly way for referencing the measurement adapter. The *supportedFormats* attribute allows knowing about the kind of formats that the adapter could translate to C-INCAMI/MIS.

When you have different data sources monitoring the same entity under analysis (e.g. different weather radars making complementary monitoring for the same region), you could grouping them under the *TraceGroup* class (See in figure 1).

The *DataSourceProperty* class (See in Figure 1) inherits from the *Attribute* class (*requirements* package in C-INCAMI) and represents each property that allows characterizing a data source (or measurement device). In this way, the *DataSourceProperties* contains the characteristics that describe to the data source, indicating the relevance and the associated value. Additionally, it incorporates a new association between *Metric* and *DataSource* called *candidates*. It represents what data sources are able for getting measures in terms of the metric definition. Moreover, through the association called *performedBy* between the *Measurement* class and *Data Source* class, the origin of the data is determined and the traceability is maintained.

Finally, the metric definition was extended through the incorporating of the concept of device's constraints by the *Constraint* class (See in figure 1). The constraints allow defining the minimum requirements that the measurement devices must satisfies before implementing the metric (e.g. minimum accuracy). A constraint is associated with one data source property but a metric may has a set of constraints linked. Each constraint incorporates the procedure for filtering (*filterAlgorithm* attribute in Figure 1) and the kind of filter (*filterType* in Figure 1. e.g. mandatory or preferable). It is possible to limit the valid values associated with a data source property by pattern (*PatternConstraint* Class in Figure 1) or explicitly indicating a set of valid values (*ConstraintValue* class in Figure 1).

Therefore, considering the constraints associated with the metrics and the available data sources, it is possible to know the candidates data sources and the data sources linked with the measurement.

This extension of the C-INCAMI framework[6] was meaningful because it allowed to consider the measures not just from the point of view quantitative but also spatial and temporal too (e.g., for the WR of the EAS INTA Anguil, we can keep the geographic data/plain text/picture/video/audio and the quantitative measures jointly for each sampling point). Moreover, it is possible to define constraints that allow being more selective in terms of the available data sources for implementing different metrics.

2.2 Case Based Organizational Memory

A key asset in any organization is related to the learned experience. In this way, the searching for getting a way which the knowledge is organizated for feedbacking the decision-making process represents a central aspect[16]. Additionally, if the knowledge could be structured as cases from the representational point of view, then it would be possible to use a case based reasoning for recommending courses of actions in each "new" situation.

Thus, the organizational memory ontology aims to be at a generic level from which other representations for specific domain applications can be formulated (see Figure 2). On the one hand, the case-based organizational memory ontology is defined at a generic organizational memory level, and on the other hand, for characterizing the cases according to the specific knowledge domain and its context [16], a domain and context ontologies should also be provided.

For the WR of INTA, on one hand, the ontology domain is associated with the agroclimate ontology (i.e. each specific concept and relationships between them), and on the other hand, the context ontology is related to M&E project definition through the new extended C-INCAMI framework (i.e. it is possible to define the entity under analysis, the attributes, the associated metrics, to obtain a likelihood distribution as quantitative measure, between others).

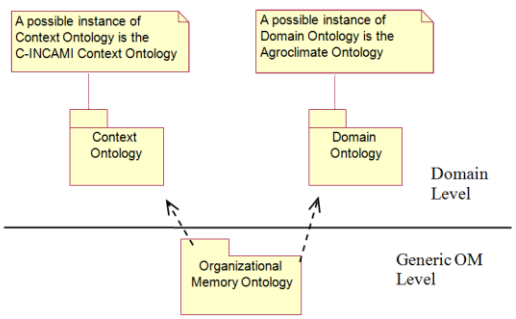

Fig. 2. The relationship between ontologies at the specific domain level and at the generic organizational memory level components

Weel now, because the organizational memory is structured in cases such as <problem, solution> pairs[6], and each problem or solution is characterized by attributes or features, the M&E project under the extended C-INCAMI associates the entity's attribute under monitor-

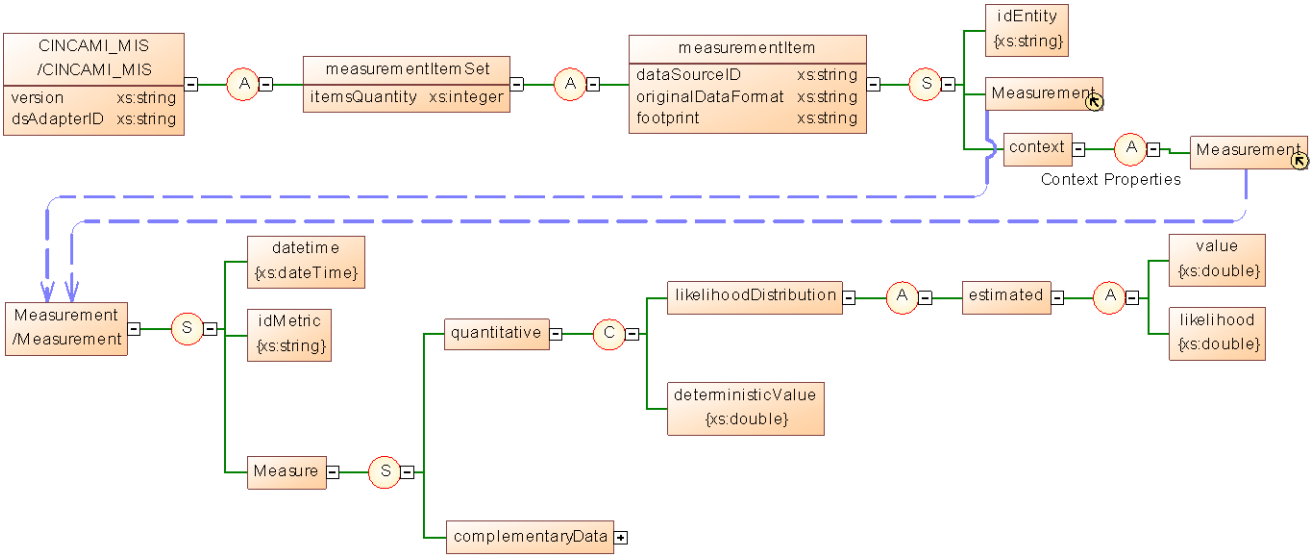

Fig.3. CINCAMI / Measurement Interchange Schema - Version 2 (A=All, S=Sequence and C=Choice).

ing with the problem's characteristic for establishing the relationships between them, serving as structured query in the case based reasoning.

3 C-INCAMI/MIS AND C-INCAMI/TS

The C-INCAMI/MIS es a measurement interchange schema based on the extended C-INCAMI ontology, which allows interchaging data and metadata jointly between heterogeneous data sources and the gathering function in PAbMM (See Section 4). The metadata are defined in the M&E projects following the GOGAME strategy, and for this reason the entity under analysis and their attributes are previosuly known. So, when each sensor is associated with a metric, it is possible to know the origin of the data and to mantain the trazability along the processing schema. In this sense, and thanks to the relationships between entity'attributes from project definition and the problem's features from the organizational memory, the recommending schema is initialized at the begining of the monitoring.

Figure 3 shows the new organization of C-INCAMI/MIS for the version 2[6]. The tag CINCAMI_MIS incorporates two attributes: *version* and *dsAdapterID*. The *version* attribute refers to the edition of the schema and the *dsAdapterID* refers to the data source adapter used for translating the measures from the raw data (e.g. from the radar) to the CINCAMI/MIS stream. Under the *measurementItemSet* tag, it is possible to have many *measurementItem* tags. Each *measurementItem* tag incorporates three attributes: 1) *dataSourceID*: it represents the identification code for the data source along the M&E projects, 2) *originalDataFormat*: the data format coming from the sensor (e.g. the weather radar) before translating through the measurement adapter, and 3) *footprint*: it allows verify the integrity of the content. In this way, using the tags *dataSourceID* and *dsAdapterID* in the CINCAMI/MIS stream, the origin of the data (e.g. the INTA Anguil Weather Radar) and the responsible for its translation (e.g. a mobile device) before the processing is known.

Under the *measurementItem* tag, you have the tags *idEntity*, *Measurement* and *context*. The *idEntity* tag identifies the entity under analysis. The *Measurement* tag describes the measurement associated with the attribute of the entity under analysis. Under the *context* tag, it is possible to group the context properties associated with the context of the entity under analysis as a set of *Measurement* tags. Therefore, you have two different *Measurement* tags but with the same internal structural organization.

The *Measurement* tag organizes the new concepts in terms of the extended C-INCAMI framework (See section 2). Here you have three tags: 1) *datetime*: it exposes the moment in that the device gets the measure, 2) *idMetric*: it indicates the metric associated with the attribute of the entity under analysis or the context property related to its context, and 3) *Measure*: it organizes the data discriminating between the *quantitative* measure and the complementary data. The *quantitative* measure can have an estimated or deterministic value. On the one hand and if the value is estimated, you will have a likelihood distribution expressed as set of <value, likelihood> pairs; and on the other hand and if the value is deterministic, you will have just one numerical value. The *complementaryData* tag is associated with the information collected at the same time that the measure (e.g. picture, video, etc) and you can get more details in [6].

However, even when C-INCAMI/MIS is very useful as measurement interchange schema, and a recommending system is present, there not exists anything for nourishing to the online classifiers with information respect to previous measures and learned experiences. It is very important because the classifiers could use the measures for the trainig and moreover, the learned experiences for improving the predictive behavior in PAbMM. In this situation emerge C-INCAMI/TS (Training Set), which establishs a new schema used at the startup of PAbMM with the aims of training the online classifiers.

Figure 4 shows the organization for the C-INCAMI/TS schema. Under CINCAMI/TS tag there are three tags: *Indicator*, *Problem* and *Solution*. The *Indicator* tag is associ-

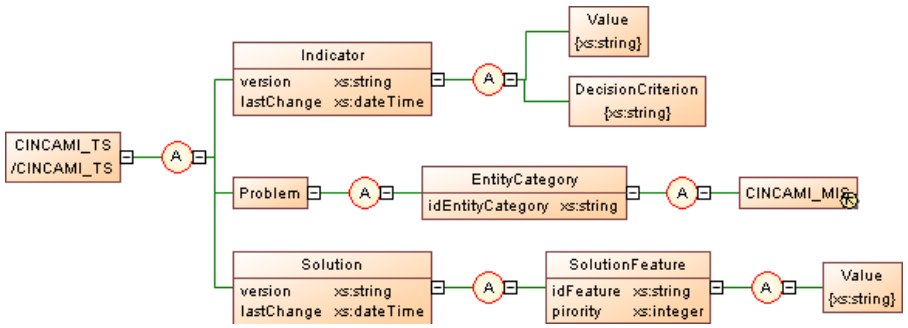

Fig.4. The C-INCAMI/TS Schema

ated with the M&E project definition (See Section 2.1), the *Solution* tag is related to the case based organizational memory (See Section 2.2), and finally, the *Problem* tag is associated jointly with the M&E project definition and the case based organizational memory.

Thus, the *Indicator* tag has two attributes: a) *version* which identifies the indicator definition in the M&E project, and b) *lastChange* which indicates the instant associated with the last updating of the indicator's definition. Under the *Indicator* tag you have the *value* tag related with the result of the interpretation of the measures from the M&E project, and the *DecisionCriterion* tag which indicates how the indicator's value was obtained from the measure's interpretation. The decision criterion could change with the time and for that reason a *version* tag is associated with the indicator. In terms of the online classifiers and their training, the indicator's value will be considered the target class.

The *Solution* tag describes the learned experiences as recommendation from the exposed situation through the characterization of the problem. Under the solution tag you have two attributes: *version* and *lastChange* which it has the same meaning with respect to the *Indicator* tag. Under the *Solution* tag, you have one or more *SolutionFeature* tags which describe a recommendation. As each recommendation's element could be given in different orders, the *SolutionFeature* tag has related two attributes: a) *idFeature*: which identifies the specific action, and b) *priority*: which establishes the order in wich the feature should be recommended. Tha *value* tag represents the quantification associated with the feature under the given priority.

On the one side, the *Problem* tag is associated with the M&E project definition, because it uses the measures that allow quantifing the attributes related to the entity under analisys and the problematic situation. On the other side, the *Problem* tag is associated with the organizational memory because the entity's attributes allow quering by characteristics the problems and get a scoring associated with the recommendation. In this sense, this entity's attributes allows describing the situation under analysis.

It is an important advance beacuse trough the CINCAMI/TS, it is possible to store the parametrized problem, the last known solution recommended by experts, and the measures that allow characterizing the problem situation in quantitative terms for better training of the online classifiers.

4 Processing Architecture based on Measurement Metadata

The PAbMM is a manager of semi-structured measurement streams, enriched with metadata supported by C-INCAMI, specialized in M&E projects, which incorporates detective and predictive behaviour at online with the ability for managing and providing large volumes of data on demand.

As shown in figure 5, the conceptual model in terms of stream processing it is as follows. The measurements are generated in the heterogeneous data sources (for example, The INTA weather radar [12]), which supply a module called Measurements Adapter (MA in Figure 5) generally embedded in measurement devices. A trace group is a set of heterogeneous data sources in where you need analyse jointly the measures.

MA incorporates together with the measured values, the metadata of the M&E project and reports it to a central gathering function (Gathering Function -GF). GF incorporates the measurem-ents streams in parallel in: a) The big data repository in persistent way, b) The C-

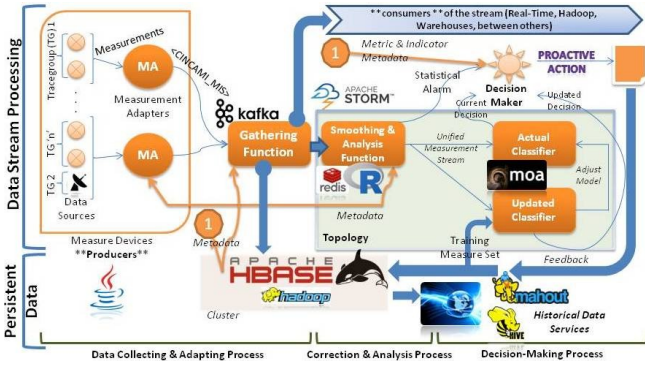

Fig. 5. Conceptual Schema of the PAbMM components

INCAMI/MIS stream for the subscribers wishing to process information at the time when it is generated (for example, for INTA weather radar data, a consumer could be the National Meteorological Service), and c) Inside a buffer organized by monitoring groups -dynamic way of grouping data sources defined by the M&E project manager- in order to allow consistent statistical analysis at level of monitoring group or by geographic region where the data sources are located, without incurring in additional processing overhead.

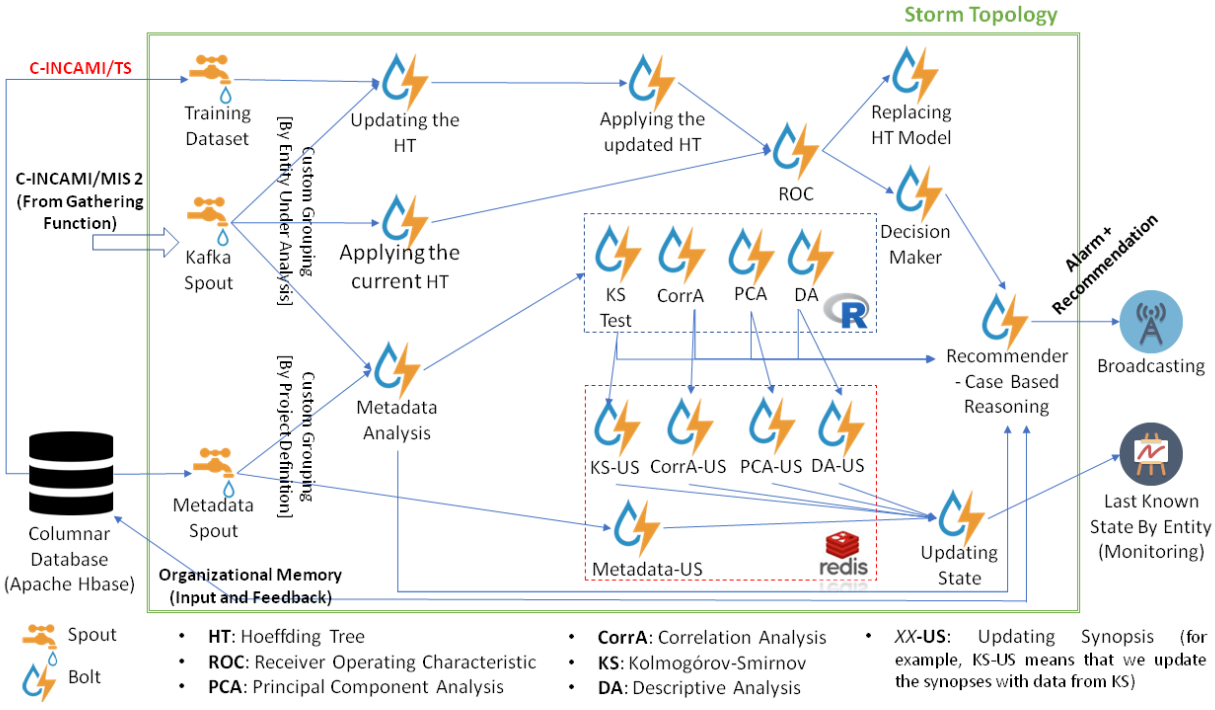

Fig.6. The C-INCAMI/TS Schema

Thus, the C-INCAMI/MIS stream, is incorporated into the big data repository with measurements and metadata, and remains available to meet requests for services associated with data on historical measurement (Big Data Repository and The Historical Data Services in Figure 1). In addition to the measurement stream is sent to the subscribed consumers, a copy of this continues within the data stream processor and applies descriptive, correlation and principal components analysis (Analysis & Smoothing Function -ASF - in Figure 1) guided by their own metadata, in order to detect inconsistent situations with respect to its formal definition, trends, correlations and / or identify system components that contribute most in terms of variability.

The shadow area associated with the storm topology in figure 5 is detailed in figure 6. In terms of Apache Storm, a Spout is source of streams inside a topology, while a Bolt is a representation for any querying, transformation and processing activity on the stream[10],[17].

In this sense and when the topology startup, a *metadata spout* (See figure 6) send the updated information associated with the M&E project definition to the follows Bolts: a) *Metadata Analysis* for monitoring the measures using the metadata on the fly, and b) *Metadata-Us*, for updating the synopses and build their initial state. In parallel, a CINCAMI/TS stream is sent trough the *Training Dataset* Spout to the *Updating the HT* Bolt. It will train initially the online classifier using the learned experiences from the organizational memory and the indicator's value as target class.

Once the storm topology has started, the C-INCAMI/MIS streams come from the GF (See figure 6) through the *Kafka Spout* and they are distributed in parallel to the follows Bolts: a) *Updating the HT*, where the hoefdding tree is updated with the new data, b) *Applying the current HT*, where the current hoefdding tree is applied and the classification from the new data is got, and c) *Metadata Analysis*, where by means of the metadata inside the C-INCAMI/MIS stream, each measure is compared with the M&E project definition for detecting eventual deviations.

On the one hand and once the Hoeffding Tree (HT) has been updated, the new tree is applied for getting a new classification. Followed, the current and updated HT are compared by a ROC (Receiver Operating Characteristic) curve[18] for knowing which model has mayor gaining. If the updated HT is better than the current HT, the updated HT will replace the current model and it will be the new current model, else the current HT will continue with its role.

On the other hand and from the metadata analysis, information about the eventual desvations is sent to the statistical analysis (KS Test, CorrA, PCA & DA in figure 6) and the recommender.

From the statistical point of view, four analyses are performed in parallel using the R Software[19] from the Storm topology: a) The Kolmogorov-Smirnov Test for knowing the distribution[20] (KS Test in figure 6), b) The correlation analysis for detecting dependencies between variables (CorrA in Figure 6), c) The Principal Component Analysis for studying the incidences of each variable in the variability of the system[21] (PCA in figure 6), and d) The descriptive analysis for characterizing the measures and the last known state [20] (DA in figure 6). With this information, the corresponding synopses grouped by trace group in Redis[22] are updated from the Storm Topology (See KS-US, CorrA-US, PCA-US & DA-US in figure 6). All the synopses will update the last known state for each trace group and entity under monitoring (See Updating State in figure 6) allowing answer queries about

the entity'state even when the streams are interrupted.

The recommender (See figure 6) receives information from: a) the decision maker (e.g. a risk situation classified following the decision criteria establishes in the M&E project definition), b) The metadata analysis bolt if an entity has associated a desviation with respect to the expert criteria in the M&E project definition, and c) the organizational memory with the similar cases. Moreover, when the recommender makes a decision based on the scoring derivated from the cases analysis, the decision is feedbacked to the organizational memory for future reasonings and sends it for broadcasting by the established media.

The storing associated with the synopsis is keeped in memory using Redis. In this way and as you can see in figure 7, a multilevel hash table for accesing and updating the values is used. Each associated value is in reality another hash table integrated for the entities that compose the trace group and you can access it using the ID of each entity. Finally, each value in the hash table of entities is an array which contains two hash tables: 1) the first hash table is organized by the ID of the metric and keeps in memory the last known measure for the metric, 2) the second hash table is organized by kind of analysis, keeping in memory the last known analysis for KS (Kolmogo-

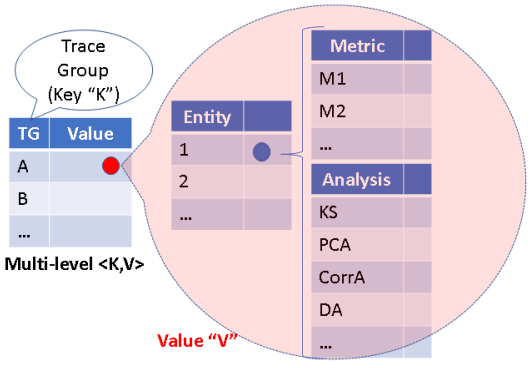

Fig. 7. Multilevel Conceptual Schema for Synopsis

rov-Smirnov), PCA (Principal Component Analysis), CorrA (Correlation Analysis) and DA (Descriptive Analysis). This new kind of organization for the synopses allows keep in memory the last know state of each tracegroup and entity for an agile answering, even allowing inform the states when the measure streams are interrupted.

5 A Practical Case: The Weather Radar

The WR are active sensors of remote sensing that emit pulses of electromagnetic energy into the atmosphere in the range of microwave frequencies. Their measurements are based, first, on the electromagnetic radiation as it propagates in the atmosphere is scattered by the objects and particles existing, and secondly, the ability of the antennas for emitting directed radiation and capturing the radiation incident from a certain direction. These sensors are tools to monitor environmental variables, and specifically, the identification, analysis, monitoring, forecasting and evaluation of hydro meteorological phenomena, as well as physical processes that these involve, given the risk that can cause severe events. The scanning region associated with the WR is dynamic and it changes its radius from 120km to 360km, considering the radar as the centre point of the circle. Additionally, the scanning happens in different angles of elevation, from 0 to 85° along the scanning territory given by the scanning radius. It is important to say that the 5° immediately upper the radar, it cannot be monitor and for that reason is called *silence cone* (See figure 8)[12].

The main applications of the WR are: a) weather description, forecasting and nowcasting, b) forecasting and monitoring of environmental contingencies (e.g. hail, torrential rain, severe storms, etc.), c) Security in navigation and air traffic control, d) Studies of atmospheric physics, e) studies of agro climatic risk, f) Provision of basic data for scientific and technological research, and g) Provision of input data for hydrological models (e.g. floods)[12].

The information recorded by the WR is collected through volumetric scans and each sampled cell has a 1 km^3 as you can see in figure 4(a). The sample units are

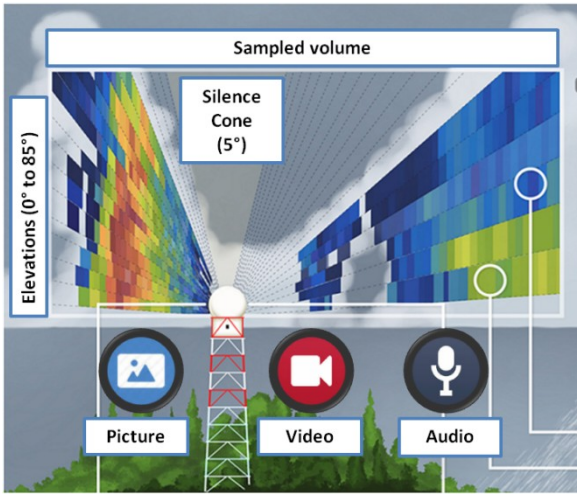

Fig. 8. Conceptual Schema Associated with C-INCAMI/MIS. The Sampled Volume represents circles around the Radar with radius from 120 km to 360 km. With C-INCAMI/MIS Version 2.0

defined as 1 km^2 and 1°. The data contains the different variables: reflectivity factor (Z), differential reflectivity (ZDR), polarimetric correlation coefficient (RhoHV), differential phase (PhiDP), specific differential phase (KDP), radial velocity (V) and spectrum width (W)[23].

From the C-INCAMI/MIS version 2 and PAbMM, important improvements are incorporated in the measurement interchanging and processing:

1. it is possible send environmental audio, descriptive picture, video of the region coming from the WR context inside the data stream,
2. it is possible verify the integrity of the multimedia data using the footprint,
3. it is feasible keep the tracebility of each measure for an given entity and data source,
4. it is feasible answer in real time with the last known state of an entity or trace group, even when the data stream could be interrupted

5. Even when the answers in the data stream processing are approximated, it is feasible complement it with the results of the analyses, for example, the correlation between metrics for an entity under analysis is useful for studing the grade of relationship between them,
6. As the measure stream processing strategy is based on a storm topology (See figure 5 and 6), it is easily updatable and extensible (e.g. for adjusting its behaviour to different kinds of WR),

As shown in figure 8, the quantitative measures (estimated or deterministic) and the complementary data (audio, video, text plain, picture & geographic information) could be associated. They allow us getting a better description of each situation under analysis in terms of the entity under monitoring and its context. In this sense, with better descriptions of each situation, it is highly likelihood improving the posterior analysis associated with the online processing strategy

6 Related Works

In this work, the perspective about PAbMM and the use of an organizational memory for recommending courses of action was given. The PAbMM uses the new schema C-INCAMI/TS for training the online classifiers in the startup, incorporating previous measurements, the entities associated, the related indicator value and the proposal of solution. Even, the storm topology associated with the statistical analysys, the synopses management, the classification and recommendation was described.

There are others processing strategy are oriented to the syntactic processing, i.e. the associated logical with the processing is embed in the application[24],[25],[26]. In the PAbMM, the M&E project definition allows guiding the processing by the use of the asociated metadata, and it allows avoiding the hard code in the logic of processing. For example, from the project definition it is possible to know the entity under analysis, the associated metrics and the interpretation of each metric's value by the indicators. Even, it is possible the use of previuos experience from the organizational memory.

In [27], an approach which conduce a semantic segmentation from a real-time sensor data stream to recognise an elderly persons complex activities is shown. PAbMM is based on a measurement and evaluation framework such as C-INCAMI, which allows getting a more comparative, repeatible and consistent measurement process. Thus, it is possible to use jointly data and metadata in PAbMM for incorporating detective and predective behaviour.

In [28] an interesting posing is associated with the use of an ontology for describing the devices under the basis of heterogeneous data sources and the data produced by them. However, PAbMM incorporates extensions to C-INCAMI framework which allows incorporating traceability and complementary data such as audio, video or picture.

7 Conclusions

In the previous work [6] we incorporate the extensions in the C-INCAMI framework to be able to manage complementary data inside each C-INCAMI/MIS stream. It was important because we could get traceability by the use of geographic information associated with the measures. In this work, the C-INCAMI/TS schema was describing, which represents an important advancing because now it is possible to train a classifier using the measures (data and metadata) and previous experience from the organizational memory. Because the organizational memory is supervised by experts in the measurement's domain through the M&E project definition, each proposed solution is based in at least one founded opinion.

Because the processing strategy is based on a storm topology, it is easily scalable and extensible. In this way, the storm topology has been detailed exposing its behaviour in predictive and detective terms. The detective behaviour is founded in the statistical analysis based on the M&E project definition, which allows catch predefined situations in real-time, for example: miscalibrations. The predictive behaviour is supported through the online classifiers, which using the previous experiences from the organizational memory they can recommend course of actions in each situation.

Finally, thanks to the incorporation of the new synopses management strategy, now it is possible incorporate jointly the statistical analysis and the last known state in memory. It is important because allows answering with the last known state associated with the entities under analysis and considering the trends, even when the data sources has been interrupted.

Acknowledgment

References

[1] W Enck, P Gilbert, S Han, V Tendulkar,B Chun,L Cox,J Jung,P McDaniel,A Sheth., "TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones," ACM Transactions on Computer Systems (TOCS), vol. 32, no. 2, p. 5, June 2014.

[2] P Rawat, K Singh, H Chaouchi, and J Bonnin, "Wireless sensor networks: a survey on recent developments and potential synergies," The Journal of Supercomputing, vol. 68, no. 1, pp. 1-48, April 2014.

[3] L. Olsina and M. Martín, "Ontology for Software Metrics and Indicators," Journal of Web Engineering (JWE), vol. 3, no. 4, pp. 262-281, 2004.

[4] H Molina and L Olsina, "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in QUATIC, Lisboa, Portugal, 2007, pp. 154–163.

[5] P. Becker, P. Lew, and L. Olsina, "Strategy to Improve Quality for Software Applications: a Process View," in International Conference on Software and Systems Process, Waikiki, Honolulu, 2011, pp. 129-138.

[6] M Diván and M Martín, "Towards a Consistent Measurement

Stream Processing From Heterogeneous Data Sources ," in International Conference on Electrical and Electronic Engineering, Johor Baru, Johor, 2017, p. In reviewing.

[7] M Diván and L Olsina, "Process View for a Data Stream Processing Strategy based on Measurement Metadata," Electronic Journal of Informatics and Operations Research, vol. 13, no. 1, pp. 16-34, June 2014.

[8] SPEM, "Software Process Engineering Meta-Model Specification," Object Management Group (OMG), Ver.2.0, 2008.

[9] M Martín and M Diván, "Case Based Organizational Memory for Processing Architecture based on Measurement Metadata," in 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2016.

[10] Apache Software Foundation. Apache Storm. [Online]. http://storm.apache.org/index.html. Last access: march 29 of 2017.

[11] Apache Software Foundation. Apache HBase. [Online]. http://hbase.apache.org/. Last acces: march 29 of 2017.

[12] M Diván, Y Bellini, M Martín, L Belmonte, G Lafuente and J Caldera, "Towards a Data Processing Architecture for the Weather Radar of the INTA Anguil," in International Workshop on Data Mining with Industrial Applications, Asunción, Paraguay, 2015.

[13] Open Geospatial Consortium and ISO, ISO 19136:2007. Geographic Information -- Geography Markup Language, 33rd ed.: International Standard Organization, 2007.

[14] Open Geospatial Consortium and ISO, ISO 19136-2:2015. Geography Markup Language (GML) -- Part 2: Extended schemas and encoding rules: ISO, 2015.

[15] International Standard Organization (ISO), ISO 639-2. Codes for the representation of names of languages -- Part 2: Alpha-3 code: International Standard Organization (ISO), 1998.

[16] M Martín, "Organizational Memory Based on Ontology and Cases for Recommendation System". PhD Thesis Computer Science School, Unviersidad Nacional de La Plata. Buenos Aires, Argentina, 2010.

[17] A. Jain and A. Nalya, Learning Storm. Create real-time stream processing applications with Apache Storm. Birmingham, United Kingdom: Packt Publishing Ltd., 2014.

[18] C Marrocco, R Duin, and F. Tortorella, "Maximizing the area under the ROC curve by pairwise feature combination," ACM Pattern Recognition, pp. 1961-1974, 2008.

[19] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing, 2017.

[20] NIST/SEMATECH, e-Handbook of Statistical Methods. USA: National Institute of Standards and Technology (NIST) - US Department of Commerce, 2013.

[21] P. Diggle, P. Heagerty, K. Liang, and S. Zeger, Analysis of Longitudinal Data, 2nd ed. Oxford (England): Oxford University Press, 2002.

[22] M Da Silva and H Tavares, Redis Essentials. Birmingham: Packt Publishing, 2015.

[23] Gematronik, Rainbow® 5 Products & Algorithms. Neuss, Germany: Gematronik GmbH, 2005.

[24] E Kalyvianaki, M Fiscato, T Salonidis, and P Pietzuch, "THEMIS: Fairness in Federated Stream Processing under Overload," in ACM SIGMOD, San Francisco, CA, USA, 2016.

[25] M Lee, M Lee, S Hur, and I Kim, "Load Adaptive and Fault Tolerant Distributed Stream Processing System for Explosive Stream Data ," Transactions on Advanced Communications Technology, vol. 5, no. 1, pp. 745-751, 2016.

[26] J Samosir, M Indrawan-Santiago, and P Haghighi, "An evaluation of data stream processing systems for data driven applications," Procedia Computer Science, vol. 80, pp. 439-449, June 2016.

[27] D Triboan, L Chen, F Chen, and Z Wang, "Semantic segmentation of real-time sensor data stream for complex activity recognition," Personal and Ubiquitous Computing, vol. 21, no. 1, pp. 1-15, February 2017.

[28] E Kharlamov et al., "Ontology-Based Integration of Streaming and Static Relational Data with Optique," in International Conference on Management of Data, New York, 2016, pp. 2109-2112